Measuring the quality of maximin space-filling designs

Francois Wahl, Cécile Mercadier, Céline Helbert

Measuring the quality of maximin space-filling designs

François Wahl [*1,2], Cécile Mercadier [†1], and Céline Helbert [‡1]

[1]Université de Lyon, CNRS, Université Lyon 1, Institut Camille Jordan, 43 blvd du 11 novembre 1918, 69622 Villeurbanne - France

[2]Institut français du pétrole, CEDI René Navarre, BP 3, 69390 Vernaison - France

March 4, 2015

Abstract

We present here a new index for measuring the quality of maximin space-filling design for computer experiments. This index is based on a very accurate approximation of the distribution of the minimum distance for uniform designs. Expressions are explicitly given in terms of closed polynomial forms for any L_p distances, including L_2, L_1 and L_∞ distances. When the size of the design or the dimension of the space is large, approximations through extreme value theory are exhibited. Some illustrations of our index are presented on simulated data and on a real problem.

Keywords : maximin, space-filling design, computer experiments, quality, extreme value theory.
AMS Classification: Primary 62K99

1 Introduction

Some of the major challenges for the automotive industry are the reduction of the greenhouse gas emissions, the fossil fuel dependency and the local pollution. In many cases, phenomenological models are not predictive enough, and these objectives rely on engine calibration which consists in determining experimentally the optimal set of parameters which satisfies some given requirements. Indeed, in order to get a usable description of the engine under study, automotive industrials launch experimental studies on test benches. Once the design is selected and the responses observed, the usual approach is to approximate the relation between the inputs and the output by means of a metamodel **?**, which can then be viewed as the physical process (see **?** and **?** for designing, modeling and analyzing physical experiments). An application of these principles is done for instance in **?** where the goal is to meet european Euro 5 emission regulations.

Since the physical model is unkown, the quality of the approximation obtained through the metamodel depends strongly on the points of simulations, which should spread points evenly throughout the experimental region to cover all the input space (**?**). This is called the *space-filling* property.

However, simple random sampling for small samples in high-dimensional regions often exhibits clustering and poorly covered areas, as remarked in **?**. In practice, LHS (namely Latin Hypercube Sampling) are very commonly performed because of their ease of use and since they guarantee that the points will be evenly spaced when projecting onto each factor separetely in the experimental region. But LHS is not necessarily *space-filling*.

[*]francois.wahl@univ-lyon1.fr
[†]mercadier@math.univ-lyon1.fr
[‡]celine.helbert@ec-lyon.fr

Thus, additional criteria are necessary to select *good* space-filling. A review can be explored in **?**. Two main kinds of criteria fit the purpose of improving the design. A first family is based on the distances between points, a design will be regarded as being of good quality if all points are far from each other. **?** have proposed two well known algorithms for distance based design:

- *maximin* distance: designs that maximize smallest distance between any two points in the design.
- *minimax* distance: designs that minimize the largest distance between any point in the experimental region and the design.

Another family of criteria, called *discrepancy* criteria, attempts to quantify the deviation from the uniform distribution. We refer for instance to **??** and **?**. A design will be considered of good quality if the empirical cumulative distribution function is close to the uniform cumulative distribution function, i.e. if its discrepancy is low.

The aim of this paper is to assess the quality of a design through the *maximin criterion*, since it is very commonly used in practice, through a normalized index of quality which will work whatever the dimension of the space (number of factors) and the size of the design (the number of points in the design). This will allow comparisons between several designs. The user will then be helped in his decision to keep the design or to generate a better one.

Though (L_2) euclidean distance remains by far the most used in practice, (L_1) Manhattan and (L_∞) Chebyshev distance and more generally (L_p) Minkowsky distance will also be considered. A first motivation comes from the form of the experimental domain that is often hypercubic and that does not correspond specifically to euclidian distance. Secondly, in the field of computer experiments, gaussian process model remains the first choice, and the covariance structure is directly linked to the distance choice. For example L_1 corresponds to Ornstein-Uhlenbeck covariance function. Finally, the covariance structure is rarely isotropic by rotation, it then gives a special role to the axes.

The paper is organized as follows. Section 2 provides some probabilistic characteristics of the distance between two random points drawn independently from the uniform distribution in the unit hypercube. It appears that this distribution can be expressed in most situations as closed polynomial form. Considering then that the distances of pairs among N points are almost independent, Section 3 state the associated approximation for the distribution of the minimum distance. Even based on some closed forms, these distributions become untractable when the dimension of the space or the size of design increases. Some approximations are developed in Section 4. Taking profit of all this theoretical material, we introduce in Section 5 a new index of the quality of a maximin space-filling design. This new measure is more instructive than a distance since, obtained from the probability measure of a particular event, it is normalized in some sense. Finally in Section 6 we illustrate the performance of the index as a measure of quality through a simulation study and on an example of space-filling design in engine calibration.

2 Distance of a pair

Let d be an integer larger or equal to one. Denote by H_d the d-dimensional hypercube $[0,1]^d$. Consider $\mathbf{X} = (X_1, \ldots, X_d)$ and $\mathbf{Y} = (Y_1, \ldots, Y_d)$ two independent random points in H_d. Both points have independent and identically distributed (i.i.d.) margins from the standard uniform distribution on $[0,1]$. Let us denote by $D_{p,d} = \left(\sum_{i=1}^{d} |X_i - Y_i|^p\right)^{1/p}$ and $D_{\infty,d} = \max_{i=1,d} |X_i - Y_i|$ the Minkowsky and the Chebyshev distance. With $p = 1$ and $p = 2$ we obtain the Manhattan and the usual Euclidean distance respectively between $\mathbf{X}$ and $\mathbf{Y}$.

Denote by $G_{p,d}$ and $g_{p,d}$ the cumulative distribution function (c.d.f.) and the probability density function (p.d.f.) of the random variable $D_{p,d}$, for $p = 1, 2, ..., \infty$. In dimension $d = 1$, the distributions are all the same with p.d.f. $g_{p,1}(x) = (2 - 2x)\mathbf{1}_{[0,1]}(x)$ and c.d.f.

$G_{p,1}(x) = (2x - x^2)\mathbf{1}_{[0,1]}(x) + \mathbf{1}_{]1,\infty]}(x)$.

The aim of the next proposition is to give exact expressions for $G_{p,1}(x)$ on $[0,1]$, for $p = 1, 2, ..., \infty$. Under the distance L_p and for general values of d, the whole support of the distribution is $[0, d^{1/p}]$. On the restricted interval $[0,1]$, expression of $G_{p,d}(x)$ can be exhibited in polynomial form.

Proposition 2.1 (**Under the distance L_p**). *The c.d.f. $G_{p,d}$ of the distance between two points from the hypercube H_d with i.i.d. uniform margins is polynomial on $[0,1]$ and satisfies the following properties.*

- *In dimension $d \geq 2$, under the L_p distance and for $x \in [0,1]$, it has the general expression*

$$G_{p,d}(x) = \sum_{\ell=d}^{2d} a_{p,d}^{(\ell)} x^\ell$$

where

$$a_{p,d}^{(\ell)} = (-1)^{l-d} \binom{d}{l-d} \left(\frac{2}{p}\right)^d \frac{\Gamma(1/p)^{2d-l}\Gamma(2/p)^{l-d}}{\Gamma(l/p+1)}.$$

- *In dimension $d \geq 1$, under the L_∞ distance and for $x \in [0,1]$,*

$$G_{\infty,d}(x) = (2x - x^2)^d.$$

- *In dimension $d = 1$ and for $x \in [0,1]$, the distribution of the distance is independent of the type of distance considered with c.d.f. $G_{p,1}(x) = (2x - x^2)$.*

- *In dimension $d \geq 2$, under the L_1 distance and for $x \in [0,1]$*

$$a_{1,d}^{(\ell)} = (-1)^{l-d} \binom{d}{l-d} \frac{2^d}{l!}.$$

- *In dimension $d \geq 2$, under the L_2 distance and for $x \in [0,1]$*

$$a_{2,d}^{(\ell)} = (-1)^{l-d} \binom{d}{l-d} \frac{\pi^{(2d-l)/2}}{\Gamma(l/2+1)}.$$

Proof. These formula can be checked by recurrence. We focus on the general formula, available for any positive value of p and any $x \in [0,1]$. Easy calculus gives $G_{p,1}(x) = 2x - x^2$.
Denoting $f_{p,d}$ the p.d.f. of the random variable $D_{p,d}^p$, we have for $d = 1$

$$f_{p,1}(x) = \frac{2}{p}\left(x^{1/p-1} - x^{2/p-1}\right)$$

and one can use the recurrence rule for higher values of d

$$f_{p,d}(x) = \int_0^x f_{p,d-1}(u) f_{p,1}(x-u) du \ .$$

The p.d.f. of $D_{p,d}$ is deduced from $g_{p,d}(x) = px^{p-1} f_{p,d}(x^p)$ and the c.d.f. as a primitive.
The formula for $a_{p,d}^{(\ell)}$ is easily checked for $d = 2$ since

$$G_{p,2}(x) = \frac{2}{p} B\left(\frac{1}{p}, \frac{1}{p}\right) x^2 - \frac{8}{3p} B\left(\frac{2}{p}, \frac{1}{p}\right) x^3 + \frac{1}{p} B\left(\frac{2}{p}, \frac{2}{p}\right) x^4 \ .$$

Assuming the formula is true until $d-1$, and applying the recurrence rule, we have to evaluate

$$f_{p,d}(x) = \sum_{\ell=d-1}^{2(d-1)} \frac{\ell}{p}\, a_{p,d-1}^{(\ell)} \int_0^x u^{\ell/p-1} f_{p,1}(x-u)du.$$

The primitive of the $u^{\ell/p-1} f_{p,1}(x-u)$ are the beta functions. Finally, we obtain

$$f_{p,d}(x) = \sum_{\ell=d-1}^{2(d-1)} \frac{2\ell}{p^2}\, a_{p,d-1}^{(\ell)} \left(x^{\frac{\ell+1}{p}-1} B\left(\frac{\ell}{p}, \frac{1}{p}\right) - x^{\frac{\ell+2}{p}-1} B\left(\frac{\ell}{p}, \frac{2}{p}\right)\right) .$$

Using the relations $B(x,y) = \dfrac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}$ and $\Gamma(x+1) = x\Gamma(x)$ and after some simplifications, we get the expected expression for $a_{p,d}^{(\ell)}$.

Comments.

- The expressions of $G_{p,d}(x)$ are generally not polynomial when $x \geq 1$. When $p = 2$, expressions of $G_{p,d}(x)$ involve trigonometric functions, as can be seen in **?** and **?** who gives detailed expressions until $d = 4$.

- Under L_1 distance, x varies on $[0,d]$, and $G_{1,d}(x)$ is piecewise polynomial on every interval $[i-1,i]$ for all i in $[1,d]$. Expressions for these polynomials can be found by recurrence from their p.d.f.. In dimension 1, we have $g_{1,1}(x) = 2(1-x)$ for $x \in [0,1]$. In dimension d and for $i \in \{2,\ldots,d-1\}$

$$g_{1,d}(x) = \begin{cases} \int_0^x g_{1,d-1}(u) g_{1,1}(x-u)du, & \text{if } x \in [0,1], \\ \int_{x-1}^x g_{1,d-1}(u) g_{1,1}(x-u)du, & \text{if } x \in [i-1,i], \\ \int_{x-1}^d g_{1,d-1}(u) g_{1,1}(x-u)du, & \text{if } x \in [d-1,d]. \end{cases}$$

  The reader should be aware that when x lies in $[i-1,i]$, the expression of $g_{1,d-1}(u)$ takes a first polynomial form on $[x-1,i-1]$ and a second on $[i-1,x]$. Even if the detailed results of this convolution are not shown here, exact calculations can be done with any symbolic manipulation software. If these softwares are not available, numerical approximated results can still be easily obtained for any L_p distances and any dimension d, by repeatedly drawing a pair of two random points and taking their distance.

- The first term $a_{p,d}^{(d)} x^d$ in $G_{p,d}(x)$ is exactly the volume of a hypersphere of radius x in dimension d.

3 Smallest distance between any pair from a design of N points

Let $\mathbf{X}^{(1)} = (X_1^{(1)},\ldots,X_d^{(1)}),\ldots,\mathbf{X}^{(N)} = (X_1^{(N)},\ldots,X_d^{(N)})$ be independent random points in the hypercube H_d, such that any margins have uniform distribution on $[0,1]$. The set of N points $\mathbf{X}^{(1)},\ldots,\mathbf{X}^{(N)}$ will represent the experimental design in Section 5 and 6.

As already explained in the introduction, from the theoretical point of view, the variable of interest is the smallest distance between any pair from the design. It has the general expression

$$\Delta_{p,N,d} := \min_{\substack{1\leq i,j\leq N \\ i\neq j}} D_{p,d}(\mathbf{X}^{(i)}, \mathbf{X}^{(j)}) .$$

Denote by $H_{p,N,d}$ and $h_{p,N,d}$ the cumulative probability function (c.d.f.) and the probability density function (p.d.f.) of the random variable $\Delta_{p,N,d}$, for $p = 1, 2, ..., \infty$.

The distances $D_{p,d}(\mathbf{X}^{(i)}, \mathbf{X}^{(j)})$ are clearly not independent since each point contributes to $N-1$ normalized distances. However, one can check that the independence assumption yields a good approximation when taking the minimum as soon as the value of N is not too small and the dimension $d \geq 2$, e.g. $N = 30$ points in dimension $d = 10$. In other words, one can observe that the smallest distance satisfies for N large enough

$$H_{p,N,d}(x) \simeq 1 - (1 - G_{p,d}(x))^{N(N-1)/2} \ . \tag{1}$$

Note that in dimension $d = 1$, it is well known from the theory of uniform spacing that $H_{p,N,1}(x) = 1 - (1 - (N-1)x)^N$ (see **?**).

The independence approximation is illustrated in Figure 1. The probability density function $h_{p,N,d}$ are obtained from simulation based on 1000 runs of the random variable $\Delta_{p,N,d}$ for the two distance L_1 (left panel) and L_2 (right). The approximations from Equation (1) are plotted in terms of density, corresponding to

$$h_{p,N,d}(x) \simeq \frac{N(N-1)}{2} g_{p,d}(x) \, (1 - G_{p,d}(x))^{N(N-1)/2-1} \ .$$

For large N, this approximation is very accurate for any value of d: the dotted curves match very closely the solid lines. This would also be the case for other distances. However when N is small, differences are large enough to be noticed.

4 Approximations from extreme value theory

As the dimension increases, highest degree of polynomials involved in the expressions of Proposition 2.1 increase as well, thus making their evaluation numerically difficult. Some approximations that may prove useful are provided for large number d of factors, or large design size N. These approximations both rely on the independence approximation. Note also that when the dimension of the input space augments, it is natural to increase the number N of points in the space-filling design as well.

Two kinds of approximation will be developed in this section, Weibull and Gumbel approximations

4.1 Weibull approximation

Weibull approximations of $H_{p,N,d}$ are valid when N the number of points is large and/or when x is close to 0. They are obtained by retaining only the first term in the expression of $G_{p,d}(x)$ and taking the limit in $H_{p,N,d}$.

Let us start by the approximation of the distribution of the distance of a pair, already discussed in Section 2.

Proposition 4.1. *As x tends to zero, $G_{p,d}(x) = c_{p,d}x^d + o(x^d)$ with $c_{2,d} = \dfrac{\pi^{d/2}}{\Gamma(d/2+1)}$, $c_{1,d} = \dfrac{2^d}{d!}$, $c_{\infty,d} = 2^d$, and more generally*

$$c_{p,d} = \left(\frac{2}{p}\right)^d \frac{\Gamma(1/p)^d}{\Gamma(d/p+1)} \ .$$

Combining the independence approximation (1) and the expansions of $G_{p,d}$ at the point 0 from Proposition 4.1 leads to

$$\begin{aligned} H_{p,N,d}\left(\frac{x}{\left(\frac{N(N-1)}{2}\right)^{1/d}}\right) &\simeq 1 - \left(1 - G_{p,d}\left(\frac{x}{\left(\frac{N(N-1)}{2}\right)^{1/d}}\right)\right)^{N(N-1)/2} \\ &\simeq 1 - \left(1 - c_{p,d}\frac{x^d}{\frac{N(N-1)}{2}}\right)^{N(N-1)/2} \xrightarrow[N\to\infty]{} \left(1 - \exp(-c_{p,d}x^d)\right) \mathbf{1}_{x>0} \ . \end{aligned}$$

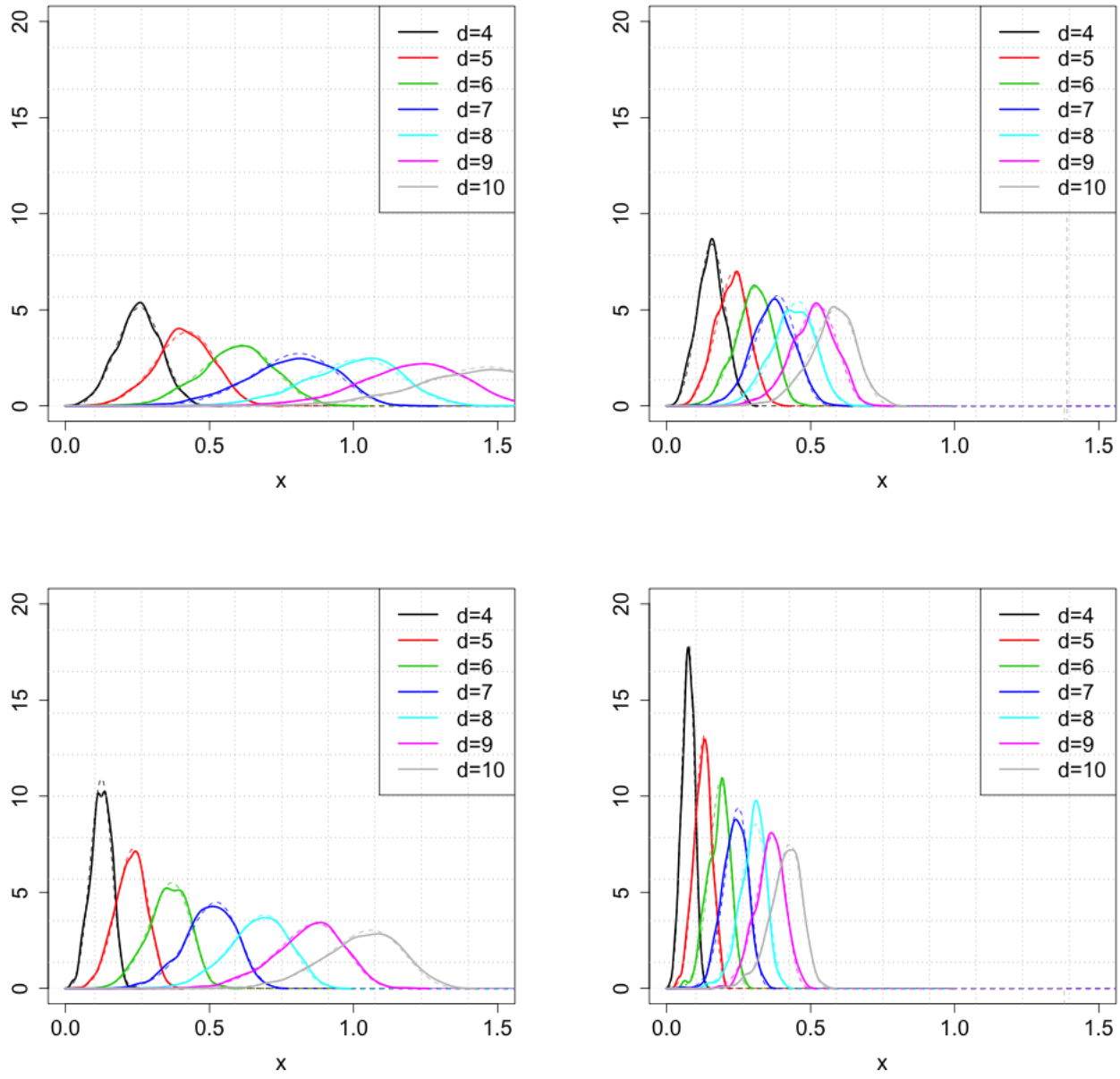

Figure 1: **Independence approximation.** Probability density function $h_{p,N,d}$ obtained from simulation (solid line) and its approximations from Equation (1) (dashed line). Distance L_1 (left) and distance L_2 (right). Number of points $N = 25$ (first line) and $N = 100$ (second line).

The limit is the well-known Weibull distribution with shape and scale parameters respectively equal to d and $c_{p,d}^{-1/d}$. Consequently, as soon as N is large enough

$$H_{p,N,d}(x) \simeq \left(1 - \exp\left(-c_{p,d}\frac{N(N-1)}{2}x^d\right)\right)\mathbf{1}_{x>0}\,. \tag{2}$$

Figure 2 plots the density $h_{p,N,d}$ of the random variable $\Delta_{p,N,d}$ for $N = 100$ and $N = 1000$, with its approximation deduced from (2). Note that these approximations are more accurate for small values of x and large values of N. However, they remain a powerful tool since they only require the knowledge of one parameter, namely $c_{p,d}$, that is known for any values of d and any L_p distances (including $p = \infty$).

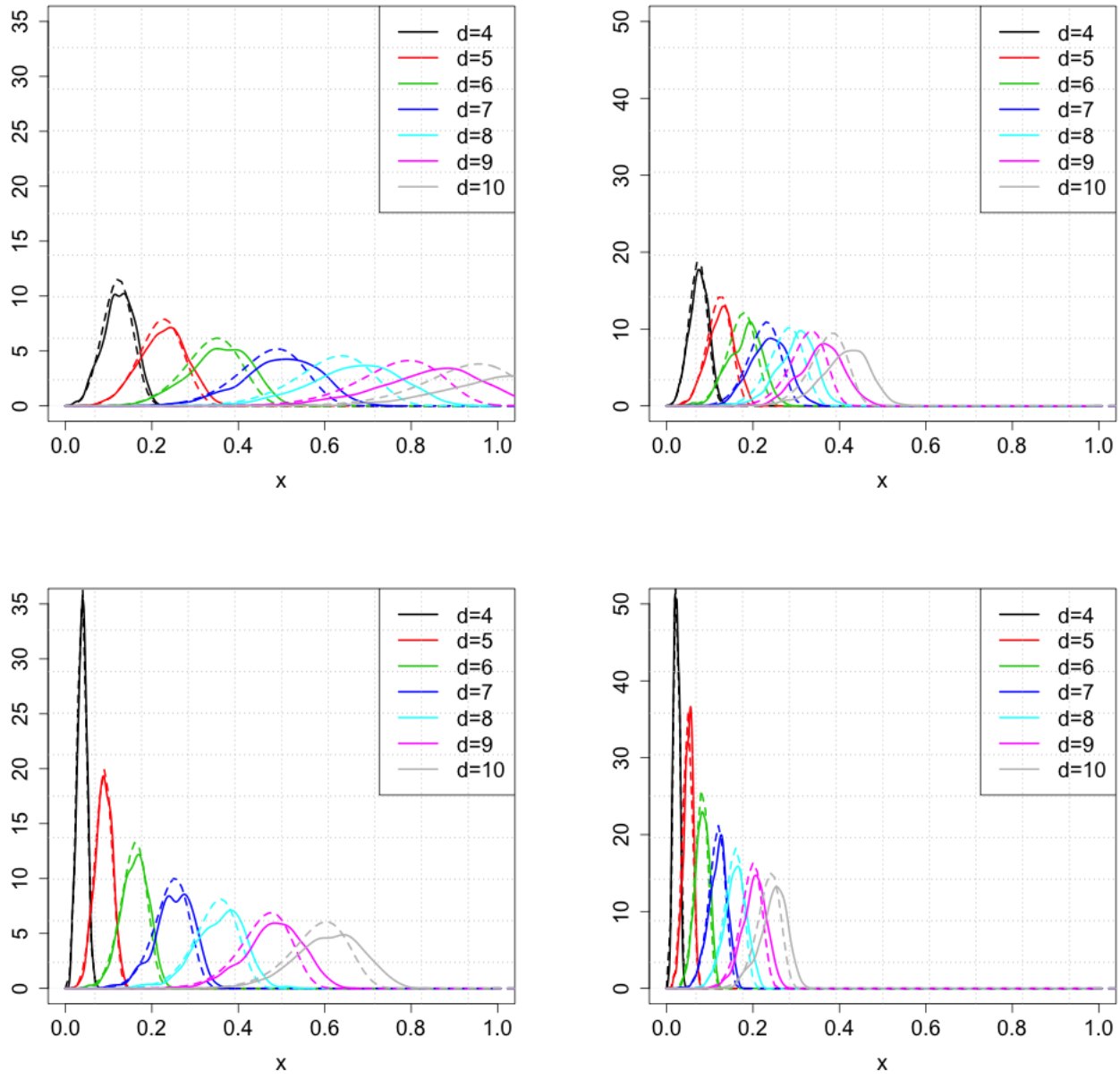

Figure 2: **Weibull approximation.** Probability density function $h_{p,N,d}$ obtained from simulation (solid line) and its approximations from Equation (2) (dashed line). Distance L_1 (left) and distance L_2 (right). Number of points $N = 100$ (first row) and $N = 1000$ (second row).

4.2 Gumbel approximation

It is also possible to propose an approximation available when the dimension d becomes very large, and this approximation is linked again to the extreme value distributions. Indeed, in a first step, as $D_{p,d}^p = \sum_{i=1}^d |X_i - Y_i|^p$, for large d, by the application of the Central Limit Theorem, we obtain a normal approximation for $D_{p,d}^p$. In a second step, we use the fact that the law of the minimum of n independent gaussian normal variables tends toward a Gumbel law.

We start by applying the Central Limit Theorem on the distribution of a pair. In the following we restrict ourself to $p = 1$ and $p = 2$ for simplicity reasons.

Proposition 4.2. *Let Φ stands for the standard normal cumulative distribution function. From*

the application of the Central limit Theorem,

$$G_{1,d}(x) \sim_{d\to\infty} \Phi\left(\frac{x - d\mu_1}{\sqrt{d}\sigma_1}\right) \tag{3}$$

with $\mu_1 = 1/3$, $\sigma_1^2 = 1/18$ *and*

$$G_{2,d}(x) \sim_{d\to\infty} \Phi\left(\frac{\frac{x^2}{d} - \mu_2}{\sigma_2/\sqrt{d}}\right) \tag{4}$$

with $\mu_2 = 1/6$ *and* $\sigma_2^2 = 7/180$.

Figure 3 displays the performance of the Gaussian approximations (3) and (4). Note that the Central limit Theorem furnishes a more accurate expansion under the distance L_1 than under the distance L_2. It is the main reason why we only present the Gumbel approximation under the distance L_1.

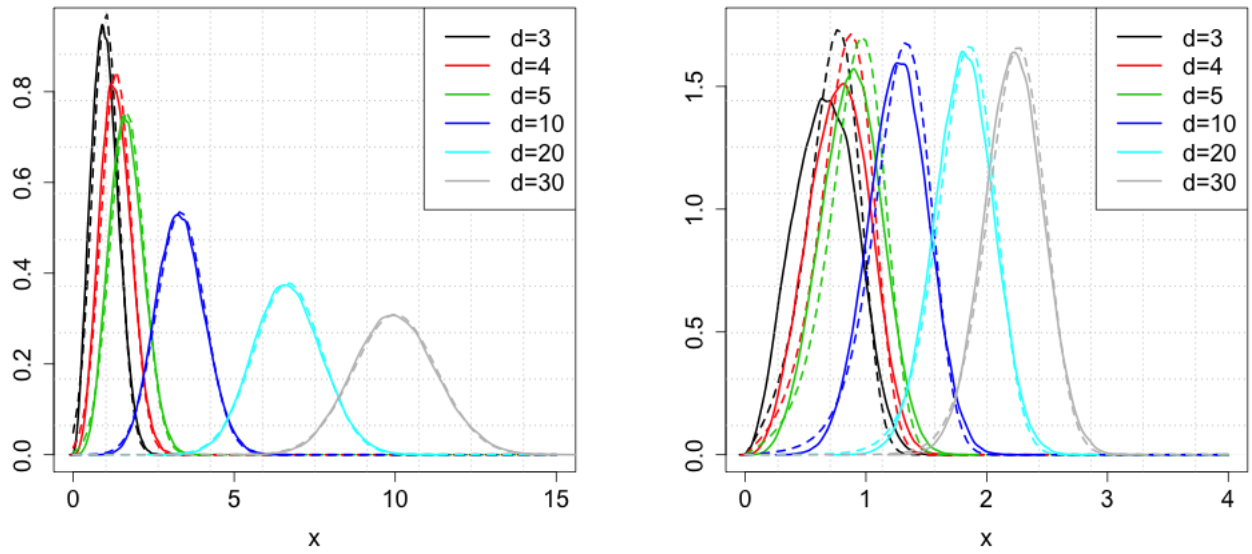

Figure 3: **TCL approximation.** Probability density function $g_{p,d}$ (solid line) and its approximations from the Central limit Theorem (dashed line). Distance L_1 (left) and distance L_2 (right).

First recall that if $Z_1, \ldots, Z_n$ are n independent and identically distributed random variables from the standard normal distribution, then $m_n = \min\{Z_1, \ldots, Z_n\}$ satisfies

$$\mathbb{P}(a_n(m_n - b_n) \leq x) \xrightarrow[n\to\infty]{} 1 - \exp(-\exp(x))$$

for $a_n = (2\log n)^{1/2}$ and $b_n = -a_n + \frac{\log\log n + \log(4\pi)}{a_n}$.
Since we know from Proposition 4.2 that for $\mu_1 = 1/3$ and $\sigma_1^2 = 1/18$ and under the Mahanttan distance

$$G_{1,d}(x) \sim_{d\to\infty} \Phi\left(\frac{x - d\mu_1}{\sqrt{d}\sigma_1}\right) .$$

It follows that

$$H_{1,N,d}(x) \simeq \mathbb{P}\left(m_{\frac{N(N-1)}{2}} \leq \frac{x - d\mu_1}{\sqrt{d}\sigma_1}\right) ,$$

and consequently that

$$H_{1,N,d}(x) \simeq_{N,d\to\infty} 1 - \exp\left[-\exp\left\{a_{\frac{N(N-1)}{2}}\left(\frac{x - d\mu_1}{\sqrt{d}\sigma_1} - b_{\frac{N(N-1)}{2}}\right)\right\}\right] . \tag{5}$$

These approximations are illustrated on Figure 4. The expansion seems really informative for high values of d.

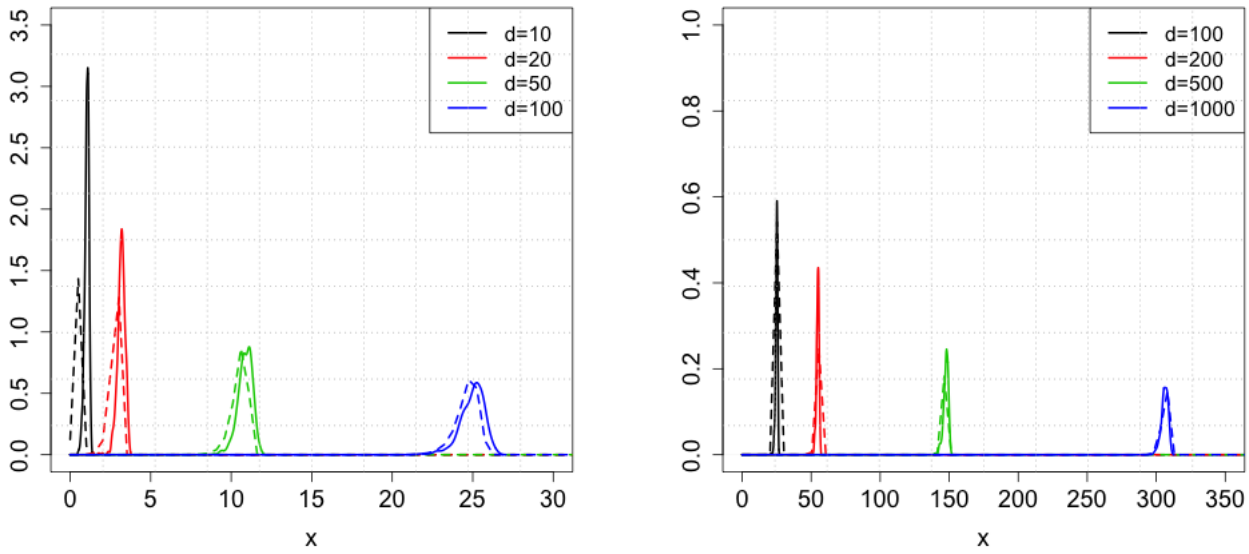

Figure 4: **Gumbel approximation.** Probability density function $h_{1,N,d}$ obtained from simulation (solid line) and its approximations from Equations (5) (dashed line). Distance L_1 only. Dimension $d \in \{10, 20, 50, 100\}$ (left) and $d \in \{100, 200, 500, 1000\}$ (right).

5 New measure for the quality of a design

Let $\mathbf{x} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ be a deterministic experimental design in a d-dimensional space. Let us denote by $\delta_{\mathbf{x}} = \delta_{\mathbf{x}}(p, N, d)$ the smallest L_p distance between any pair from $\mathbf{x}$. Taking into account the motivation of space-filling methodology, the larger the value $\delta_{\mathbf{x}}$, the better the design. From what precedes, it is possible to define the following quantity $-\log_{10}(1 - H_{p,N,d}(\delta_{\mathbf{x}}))$ as a measure of the quality of the design.

Since the distribution $H_{p,N,d}$ is not completely known, we first use the independence approximation that has been illustrated in Section 3. Let us set the index as

$$\mathbb{I}_{p,N,d}(\mathbf{x}) := -\frac{N(N-1)}{2}\log_{10}(1 - G_{p,d}(\delta_{\mathbf{x}})) .$$

Rephrased in an other way, we see that $G_{p,d}(\delta_{\mathbf{x}}) = 1 - 10^{-\mathbb{I}_{p,N,d}(\delta_{\mathbf{x}})}$, and the index measures the closeness to 1: logarithmic expression is preferred since in many cases in maximin space-filling design $H_{p,N,d}(\delta_{\mathbf{x}})$ is very close to 1, while, for small values of $\delta_{\mathbf{x}}$, the index is still calculable, in the same way.

$\mathbb{I}_{p,N,d}$ can vary from 0 to $+\infty$. An index close to 0 indicates that the minimum distance between the points of the design is easily reachable by a random uniform design, while a high value (say greater than 2) shows that the corresponding design is unlikely to come from a random uniform design, and may be resulting from an optimization process. Naturally, in maximin design, high values of $\mathbb{I}_{p,N,d}$ should be preferred.

The maximin distance for a design is bounded (**?**) and depends on both the number of points N and the dimension d. Moreover, this distance relies strongly on the type of design and the algorithm and is difficult to interpret, contrary to the proposed index, which is normalized in a certain way. Note however that the index depends on the type of distance used, so that indices for L_1 distance will not be equivalent to indices under L_2 or L_∞.

When the size of the design N is large, this definition could be replaced by

$$\tilde{\mathbb{I}}_{p,N,d}(\mathbf{x}) := \frac{(\delta_{\mathbf{x}})^d N(N-1)c_{p,d}}{2\ln(10)} ,$$

where $c_{p,d}$ is given in Proposition 4.1 . This last expression comes directly from the main idea of the index and the expansions (2).

Typical random sampling gives the curves displayed on Figure 5 where the dimension of the space varies between 2 and 10. As already noticed, the minimum distance increases with the dimension d. At the end of each curve, when the distance is sufficiently large, the c.d.f. is numerically equal to one, but the index remains calculable.

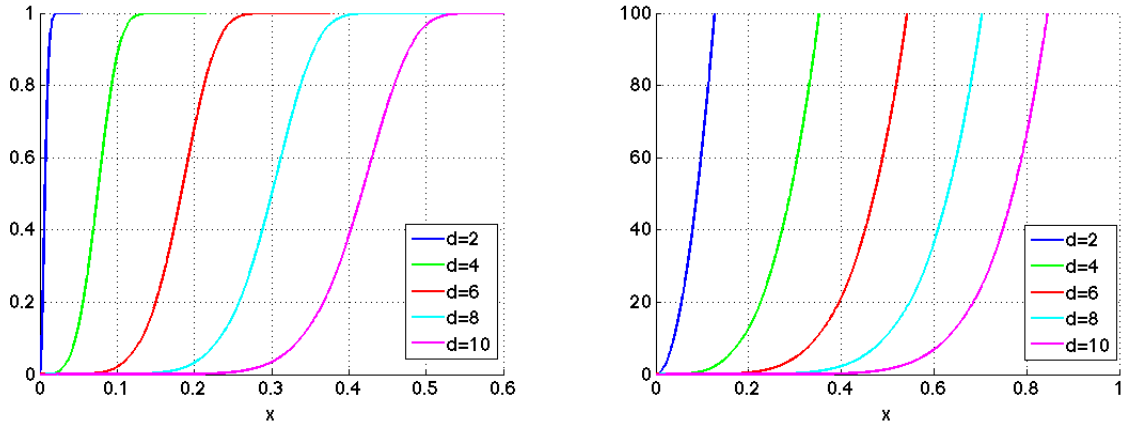

Figure 5: **Index representation under uniform designs.** Several c.d.f. $H_{2,100,d}$ (left panel) and corresponding index $\mathbb{I}_{2,100,d}$ (right panel) for $N = 100$ points drawn uniformly in the unit hypercube when the dimension d varies from 2 to 10.

One could be interested by the behaviour of our index on more reallistic designs. Indeed, the previous example was based on classical Monte-Carlo sampling. The same exercise has been continued with Latin Hypercube Sampling, a well known method since **?**. The initial output of this algorithm can be improved by running a fixed number of iterations and selecting for example the best response which maximizes the minimum distance between the points of the design (see for example *maximinlhs* from package *lhs* in R).

However, special algorithms have been proposed, based on simulated annealing algorithm, to improve a first initial drawing (**?**), for instance *maximinSA_LHS* in the package *DiceDesign* in R. In this case, the improvement clearly shows a big difference : for example, in dimension $d = 10$ with a design containing $N = 100$ points and after 1000 iterations, the mean minimum distance with the optimized routine *maximinSA_LHS* is around 0.9074, which corresponds to an index of 166, while with *maximinlhs* from package *lhs* the mean minimum distance is around 0.5568 with an index of 3.7. The results are presented in Figure 6 showing that the optimized algorithm is much more efficient than classical LHS.

This previous comparison can be further investigated for other dimensions and numbers of points. Indeed, in many applications, one is not interested in finding the "best" design but only a "good" design, with a given index *id*. In this case the number of iterations is not fixed in advance but set by the algorithms which are run until the searched *id* is reached. In this

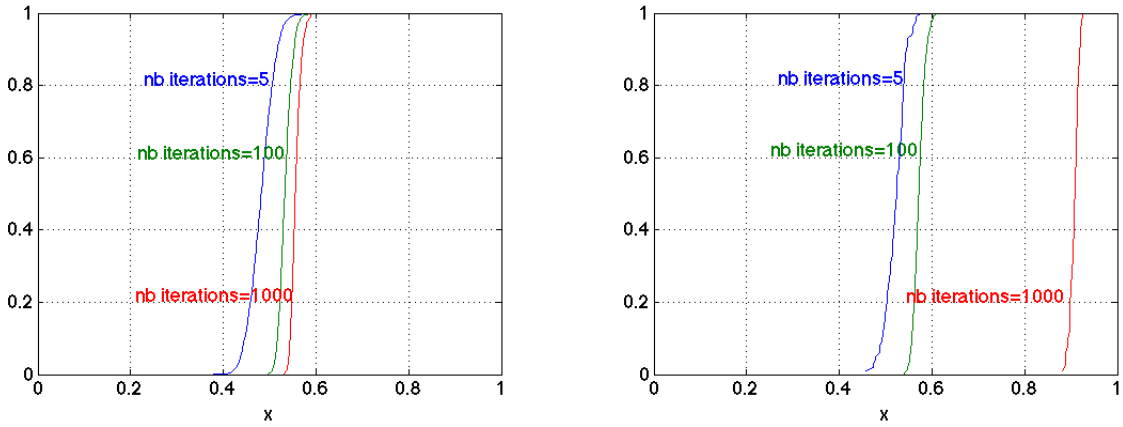

Figure 6: **Index representation under LHS.** On the left panel, we have repeated 1000 times the following procedure: we draw 5, 100 and 1000 'classical' LHS of $N = 100$ points in dimension $d = 10$ (run with *lhsdesign* from Matlab), and among them, we choose the design with the maximum minimum distance. The right panel is based on a simulated annealing optimized algorithm implemented with *maximinSA_LHS* from *DiceDesign* with the same number of iterations. The abscissa is the minimum distance between the points of the design, while the y axis is the simulated cdf for the algorithms we are considering.

context, two methods for building such a design are possible : the first consists in applying an optimization routine, the second consists in generating uniform designs until index exceeds id. The question is then: how many uniform designs should be generated before observing one with index id, compared to the number of iterations needed by the optimization routine to obtain the same result? Note that we call "uniform design", a design which points have been randomly chosen.

Let us consider a design of N points in dimension d with index id. Let δ_{id} be the minimal distance between two points of this design. From the definition of the index, δ_{id} can be approximated by $\delta_{id} = G_{p,d}^{-1}(1 - 10^{-id})$ which means that if a design has an index higher than id, its minimal distance between two points is higher than δ_{id}. In the following, we compare K, the number of uniform designs that must be generated until observing one with a maximin criterion higher than δ_{id} to R, the number of iterations of an optimization routine to produce such a design. The chosen routine is *maximinSA_LHS* from the package *DiceDesign*. This routine has been slightly adapted to produce non LHS designs, in order not to bias the comparison.

The theoretical probability distribution of K is a geometric law of parameter 10^{-id}. The mean of the distribution is then 10^{id} and the quantile of order $1 - \alpha$ is $K_{1-\alpha} = \frac{ln(\alpha)}{ln(1-10^{-id})}$, such that the probability of waiting more than $K_{1-\alpha}$ simulations before observing a design of index id is α. Note that this probability is independent of the dimension d and the number of points of the design N.

If we now want to compare the distribution of R to the distribution of K it is sufficient to compare empirical realizations of R to the theoretical law of K. Tests have been done for three values of the index ($id = 1$, $id = 3$ and $id = 5$) and two different numbers of points ($N = 100$ and $N = 500$). On the top of Figure 7, i.e. $id = 1$, it can be seen that the geometrical law performs better than the distribution of R for all the dimensions and for the two numbers of points. In this case, the expectation of K is 10 whereas the number of iterations of the optimization routine is distributed between 0 and 500 for $N = 100$ (or between 0 and 1500 for $N = 500$). When id equals 3 comparison reverses : the cumulative distribution function of R is going to 1 faster

that the geometric distribution regardless of the dimension or the number of points. Note that the comparison is less favorable when the dimension or the number of points is high, that is to say when the optimization problem is more complex. When $id = 5$, the use of an optimization method is inescapable, the expectation of the corresponding law for K being extremely high ($E(K) = \frac{1}{1-10^{-5}} = 100000$). This case is not represented on Figure 7.

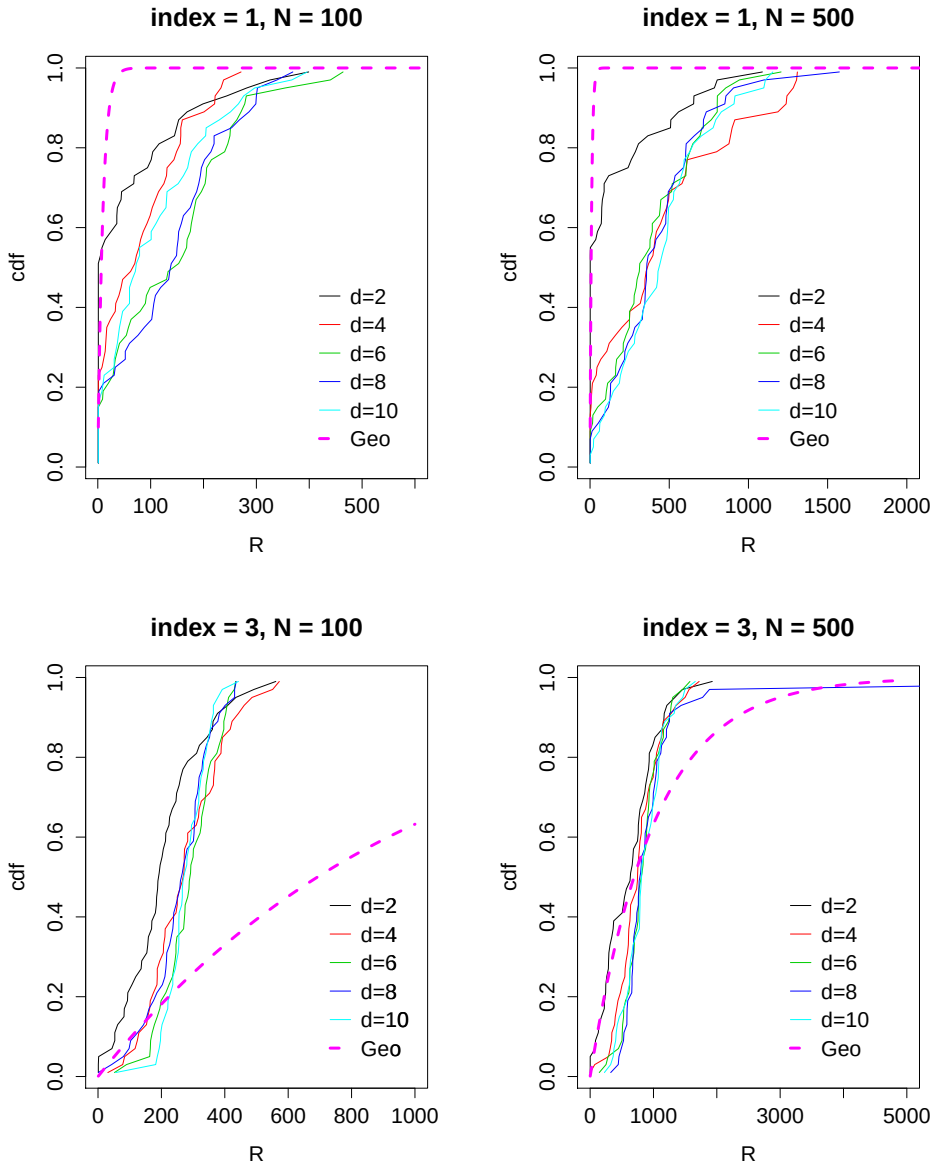

Figure 7: **Random versus optimization**. Cumulative distribution functions of R (solid line) for different dimensions $d \in \{2, 4, 6, 8, 10\}$, $id = 1$ (first row) and $id = 3$ (second row), $N = 100$ (left) and $N = 500$ (right). Cumulative distribution function of K (dashed line).

In conclusion to this comparison, we can say that if the requested quality of the space filling design is not too high, it is more efficient to independently generate uniform designs as necessary than to run an optimization procedure, especially when the optimization problem is difficult (high dimension or high number of points). Rather, an optimization routine should be preferred when the required quality is high.

Our new index is effective for design comparison whatever dimension and number of points.

Moreover, since it is normalized, when considering a standalone design, our index gives an evaluation of the quality of the design without the need of interpreting distances.

6 Application

As already indicated in the introduction, we are motivated by a real example from automotive industry. Engine calibration consists in defining the optimal tuning of parameters used by engine control strategies, for example for achieving low consumption or low pollution. Due to the increasing number of these parameters, manual tuning of engine parameters is now replaced by mathematically assisted calibration process, that is based on design of experiments. The dimension of the input space for parameters is $d = 11$, among which we can find injection engine speed, load, air flow rate, injection parameters, as described in **?**. To avoid useless complications, all the parameters are supposed to vary between 0 and 1.

To prevent the engine from going in forbidden regions, some constraints (linear and non linear) have to be added, and the final experimental area is not any longer hypercubic. Design in constrained domains is still an active domain of research. **?** and more recently **?** propose performing strategies in constrained regions based on simulated algorithm. **?** consider the case where linear constraints passing through the origin restrict the available domain.

Here we propose the following strategy to calculate the index. If we were calculating the volume of the constrained region by Monte-Carlo simulations, we would simply count the number N of points falling in the targeted region out of a total number of draws N_{tot}, and the volume would be $V = N/N_{tot}$. Conversely, N points uniformly drawn in the experimental area should correspond to N/V points in the unit hypercube where V is the volume of the constrained region. Then if $\delta_{\mathbf{x}}$ is the minimum distance obtained for the points in the experimental region, the probability that a random uniform design will have a minimum distance bigger than $\delta_{\mathbf{x}}$ in the unit hypercube is given by

$$P(\Delta_{p,N/V,d} > \delta_{\mathbf{x}}) = 1 - H_{p,N/V,d}(\delta_{\mathbf{x}}) = (1 - G_{p,d}(\delta_{\mathbf{x}}))^{\frac{N/V(N/V-1)}{2}}$$

where $G_{p,d}(x)$ is defined as in Section 3.

Under L_2 distance. Simulation of designs from the uniform distribution and selection of the points falling in the targeted region allows to estimate the volume of the constrained region around $V = 0.23$. With $N = 250$ points, the minimum L_2 distance between pairs of such design is around $\delta_{\mathbf{x}} = 0.33$, corresponding to an index of 1.06 calculated with an equivalent number equal to N/V points. As it is done n **?** or **?**, this first initial design can be improved by exchanging coordinates. After the application of this procedure, the minimum distance becomes $\delta_{\mathbf{x}} = 0.395$ and the index equals 6. As the cumulative distribution function is very stiff, this small change in the distance has an important impact on the index.

Under L_1 distance. The same calculations can be done with L_1 distances for maximin designed as in **?**. We find $\delta_{\mathbf{x}} = 0.52$ in the initial design and 0.98 after exchanging coordinates. These values correspond to an index of 0.006 for the initial design and 4 for the modified one. Note that these calculations are only indicative here, since maximin designs depend on the chosen distance.

Conclusion. In this paper, we propose a new index to measure the quality of space-filling designs, based on the probability distribution of the minimum distance of N points drawn uniformly in the unit hypercube of dimension d. This index depends on the chosen distance, the number of points and the dimension of the space. In most common situations, it can be very well approximated by polynomial form. When the number of points in the design is large or under high dimension, since formulas can become cumbersome to evaluate, we give some powerfull approximations of the index.

Acknowledgements

This work was supported by the LABEX MILYON (ANR-10-LABX-0070) of UniversitÈ de Lyon, within the program "Investissements d'Avenir" (ANR-11-IDEX-0007) operated by the French National Research Agency (ANR).

References

Y. Auffray, P. Barbillon, and J.M. Marin. Constrained maximin designs for computer experiments. *Statistics and Computing*, 22:703–712, 2012.

S. Coles. *an introduction to statistical modeling of extreme values*. Springer, 2001.

G. Damblin, M. Couplet, and B. Iooss. Numerical studies of space filling designs: optimization of latin hypercube samples and subprojection properties. *Journal of Simulation*, 7:276–289, 2013a.

G. Damblin, M. Couplet, and B. Iooss. Numerical studies of space-filling designs: optimization of latin hypercube samples and subprojection properties. *Journal of Simulation*, 7:276–289, 2013b.

H. David. *Order Statistics*. Wiley, 1980.

K. Fang, R. Li, and A. Sudjianto. *Design and Modeling for Computer Experiments*. Chapman & Hall, 2005.

J. Franco, D. Dupuy, O. Roustant, G. Damblin, and Iooss B. Package DiceDesign, 2013.

F. A. Hickernell. Generalized discrepancy and quadrature error bound. *Mathematics of computation*, 67:299–322, 1998.

B.G.M. Husslage. *Maximin designs for computer experiments*. PhD thesis, Universiteit van Tilburg, 2006.

R. Jin, W. Chen, and A. Sudjianto. An efficient algorithm for constructing optimal design of computer experiments. *Journal of Statistical Planning and Inference*, 134:268–287, 2005.

M.E. Johnson, L.M. Moore, and D. Ylvisaker. Minimax and maximin distance designs. *J. of Statistical Planning and Inference*, 26:131–148, 1990.

H. Langouet, L. Metivier, D. Sinoquet, and Q.H. Tran. Optimization for engine calibration. In *EngOpt 2008 - International Conference on Engineering Optimization*, Rio de Janeiro, Brazil, 2008.

S. Magand, L. Pidol, F. Chaudoye, D. Sinoquet, F. Wahl, M. Castagne, and B. Lecointe. Use of Etahnol/diesel Blend and Advanced Calibration Methods to Satisfy Euro 5 Emiossion Standards without a DPF. *Oil and Gas Science and Technology*, 66:855–875, 2011.

M. D. Mckay, R. J. Beckman, and W. J. Conover. Constrained maximin designs for computer experiments. *Technometrics*, 21:239–245, 1979.

H. Niederreiter. Low-Discrepancy and Low-Dispersion Sequences. *Journal of number theory*, 30:51–70, 1987.

M. Petelet, B. Iooss, O. Asserin, and A. Loredo. Latin hypercube sampling with inequality constraints. *Advances in Statistical Analysis*, 3:11–21, 2010.

J. Philip. The probability distribution of the distance between two random points in a box. *TRITA MAT MA*, 7 (10), 2007.

J. Philip. The distance between two random points in a 4- and 5-cube. *Journal of Chemometrics*, 21(5-6):198–207, 2010.

L. Pronzato and W. G. Müller. Design of computer experiments: space filling and beyond. *Statistics and Computing*, 23:681–701, 2012.

J. Sacks, W.J. Welch, W.J. Mitchell, and H.P. Wynn. Design and analysis of computer experiments. *Statistical Science*, 4:409–42, 1989.

T.J. Santner, B.J. Williams, and W.I. Notz. *The Design and Analysis of Computer Experiments*. Springer, 2003.

E. Stinstra, D. den Hertog, P. Stehouwer, and A. Vestjens. Constrained maximin designs for computer experiments. *Operations Research*, 57:595–608, 2003.

E. Van Dam, G. Rennen, and B. Husslage. Bounds for maximin latin hypercube designs. *Operations Research*, 57:595–608, 2009.